




## ARTICLE

# Item response theory in early phase clinical trials: Utilization of a reference model to analyze the Montgomery-Åsberg Depression Rating Scale

Marije E. Otto<sup>1,2</sup>  | Kirsten R. Bergmann<sup>1</sup> | Marieke L. de Kam<sup>1</sup> | Kasper Recourt<sup>1</sup> | Gabriël E. Jacobs<sup>1,3</sup>  | Michiel J. van Esdonk<sup>1</sup> 

<sup>1</sup>Centre for Human Drug Research (CHDR), Leiden, The Netherlands

<sup>2</sup>Leiden Academic Centre for Drug Research (LACDR), Leiden University, Leiden, The Netherlands

<sup>3</sup>Department of Psychiatry, Leiden University Medical Center (LUMC), Leiden, The Netherlands

## Correspondence

Marije E. Otto, Centre for Human Drug Research (CHDR), Zernikedreef 8, 2333CL Leiden, The Netherlands.  
Email: [motto@chdr.nl](mailto:motto@chdr.nl)

## Funding information

Sumitomo Pharma Co., Ltd.

## Abstract

In clinical trials, Montgomery-Åsberg Depression Rating Scale (MADRS) questionnaire data are added up to total scores before analysis, assuming equal contribution of each separate question. Item Response Theory (IRT)-based analysis avoids this by using individual question responses to determine the latent variable ( $\psi$ ), which represents a measure of depression severity. However, utilization of IRT in early phase trials remains difficult, because large datasets are needed to develop IRT models. Therefore, we aimed to evaluate the application and assumptions of a reference IRT model for analysis of an early phase trial. A cross-over, placebo-controlled study investigating the effect of intravenous racemic ketamine on MADRS scores in patients with treatment-resistant major depressive disorder was used as a case study. One hundred forty-seven MADRS responses were measured in 17 patients at five timepoints (predose to 2 weeks after dosing). Two reference IRT models based on different patient populations were selected from literature and used to determine  $\psi$ , while testing multiple approaches regarding assumed data distribution. Use of  $\psi$  versus total score to determine treatment effect was compared through linear mixed model analysis. Results showed that determined  $\psi$  values did not differ significantly between assumed distributions, but were significantly different when changing reference IRT model. Estimated treatment effect size was not significantly affected by chosen approach nor reference population. Finally, increased precision to determine treatment effect was achieved by using IRT versus total scores. This demonstrates the usefulness of reference IRT model application for analysis of questionnaire data in early phase clinical trials.

## Study Highlights

### WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

In clinical trial analyses, individual item scores in rater-based questionnaires, such as the Montgomery-Åsberg Depression Rating Scale (MADRS) are added

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *CPT: Pharmacometrics & Systems Pharmacology* published by Wiley Periodicals LLC on behalf of American Society for Clinical Pharmacology and Therapeutics.

up to total scores, assuming equal contribution of each question. Item response theory (IRT) based analysis uses item level data and has shown to result in better estimates of treatment effect. However, early phase trial data are too sparse for IRT model development.

#### **WHAT QUESTION DID THIS STUDY ADDRESS?**

Can a reference IRT model be applied to analyze MADRS data from a small clinical trial investigating ketamine's effects in patients with major depressive disorder, and are results influenced by assumptions regarding data distribution or reference population?

#### **WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?**

IRT-based analysis of ketamine's treatment effect on the MADRS was unaffected by assumed distribution or choice of reference population and yielded improved precision compared to total score analysis.

#### **HOW MIGHT THIS CHANGE DRUG DISCOVERY, DEVELOPMENT, AND/OR THERAPEUTICS?**

Use of reference IRT models for exploratory analysis of rater-based questionnaire data in early phase clinical trials may improve decision making in subsequent development phases of novel compounds.

## **INTRODUCTION**

Early proof-of-concept (PoC) clinical trials investigating novel pharmacological treatments for mental disorders, such as major depressive disorder (MDD), often utilize interview-based rating instruments as a primary end point because established biomarkers are lacking.<sup>1</sup> Currently, the Montgomery-Åsberg Depression Rating Scale (MADRS), among others, is considered one of the golden standards to determine MDD severity and to quantify treatment effects.<sup>1,2</sup> It consists of 10 items, each rated between 0 and 6 by a trained clinician during a clinical interview. The separate scores of each individual item are typically combined in a total score that monitors overall (change in) MDD severity, where higher total scores represent greater MDD severity, and vice versa, following therapeutic interventions. By doing so, it is assumed that each item contributes equally to the total score and therefore MDD severity because no item-specific weighting is applied to, for instance, core symptoms of depression as opposed to additional symptoms related to depression. For example, an individual who scores high on suicidal thoughts, but low on reduced appetite can theoretically be regarded equally depressed as an individual who scored high on reduced appetite and low on suicidal thoughts, even though the individual MDD severity might differ between such subjects. In other words, different sets of individual answers may have similar total scores, from which it could be concluded there is no difference in disease severity whereas this might not necessarily be the case from a phenomenological perspective.

In contrast to a total score analysis, item response theory (IRT) incorporates all available information from the individual questionnaire item responses into one latent variable ( $\psi$ ) value, which represents the disease severity.<sup>3-5</sup> An IRT model is based on the probabilities for answers or scores to be given to items dependent on the underlying disease severity or  $\psi$  value, also called the item characteristic curves (ICCs). The ICC's are generally defined with logit functions, which rely on two parameters: item discrimination (i.e., the ability of an item to discriminate between  $\psi$  values, or the steepness of the curve) and difficulty (i.e., the value of  $\psi$  where the probability of a certain score is 50%, or the location of the curve on the  $\psi$  scale). For the MADRS, in case of the previous example, this can be translated to higher discrimination and lower difficulty parameters for the question about suicidal thoughts compared to reduced appetite: a change in score from 2 to 3 for suicidal thoughts is a more reliable indication of increased MDD severity (item discrimination) compared to a similar increase for reduced appetite, whereas higher scores may be reached less quickly for suicidal thoughts than for reduced appetite with the same MDD severity (item difficulty).

The benefit of IRT-based analysis versus total score analysis of clinical trial data has already been demonstrated by studies using both empirical and simulation data to show that estimation of the treatment effect is less biased when applying IRT.<sup>6,7</sup> IRT has also increasingly been applied in the field of nonlinear mixed effects (NLME) modeling, where longitudinal models describing disease progression and treatment effect are now

developed based on the change in  $\psi$  over time instead of the total score.<sup>3,8–14</sup> Many of these studies have shown that the statistical power to determine treatment effect using an IRT approach is higher when compared to using the total score data in NLME modeling.<sup>9,10,13</sup> These studies either used hypothetical questionnaires for simulations or were based on data from questionnaires used in clinical practice, such as the State Trait Anxiety Index Dutch Y-version (STAI-DY), the Exacerbation of Chronic Pulmonary Disease Tool (EXACT), the Unified Parkinsons Disease Rating Scale (UPDRS), and Hamilton depression rating scale (HAMD-17). Similar benefits of IRT-based analysis are to be expected for determination of treatment effect in clinical trials using the MADRS as endpoint for MDD, however, this is yet to be demonstrated.

Unfortunately, large datasets are required to develop IRT models. Depending on the number of questions and possible categories in case of ordinal data, a relatively high number of parameters needs to be estimated. Several simulation studies have investigated this interplay of available data and parameter recovery precision, however, only few provide concrete guidelines.<sup>6,7,15</sup> For instance, it has been reported that at least 250–500 calibration measurements are necessary for accurate recovery of item parameters for a graded response model with up to 5 categories.<sup>3,16</sup> For an example study where the MADRS, with greater than five categories, is measured five times during both active and placebo treatment visits in a PoC trial of 20 subjects, this minimum number of measurements would already not be achieved. The data availability is therefore a limiting factor for the application of IRT in early phase clinical trials, yet the results from PoC trials significantly impact go/no-go decisions in subsequent development phases.

An alternative to de novo development of IRT models for PoC trials would be to apply existing reference models.<sup>3</sup> Assuming that the selected rater-based questionnaire displays similar psychometric characteristics in both the population of interest and the reference population (i.e., the population used for the IRT model development), item-specific score data can be transformed to  $\psi$  data. Such data can then be used for NLME modeling or, in case of sparse longitudinal data, statistical analyses such as linear mixed effects models or mixed models for repeated measures.<sup>6,7</sup> Still, the application of reference IRT models to investigate treatment effects in early phase clinical studies still finds itself in its infancy.

Therefore, the aim of this study is to apply a reference IRT model to a relatively small but robust clinical dataset, to (1) demonstrate the benefits of IRT in early phase clinical trials in psychiatry and (2) to assess the risk of bias related to model assumption(s). To achieve these objectives, data from an early phase clinical trial investigating

the effect of ketamine as treatment for treatment-resistant MDD using the MADRS as primary outcome were used as a case study.

## METHODS

### Clinical trial data

Data from a randomized, double-blind, placebo-controlled, cross-over study evaluating the effects of ketamine on resting state functional brain connectivity in patients with MDD, conducted at the Centre for Human Drug Research (The Netherlands) were used for this analysis. The study (EudraCT: 2016-003999-51) was executed according to the International Committee on Harmonization of Good Clinical Practice guidelines and principles of the Declaration of Helsinki and was approved by an independent ethics committee (Stichting Beoordeling Ethiek Biomedisch Onderzoek). Oral and written informed consent was obtained from the subjects before the start of the study.

Men and women between 18 and 65 years of age, who met the Diagnostic and Statistical Manual of Mental Disorders (4th Edition) diagnostic criteria for MDD without psychotic features at screening (confirmed by the Mini-International Neuropsychiatric Interview), with a HAMD-17 total score of greater than or equal to 18 at screening and first visit, and who demonstrated partial or no response to treatment with a serotonin reuptake inhibitor or serotonin-noradrenalin reuptake inhibitor despite a therapeutic dose for at least 4 weeks of treatment, were included. Use of concomitant antidepressant treatment was allowed if not newly started during the study period and if the dose remained unchanged.

Patients received a single dose of 0.5 mg/kg racemic ketamine or placebo via an intravenous (i.v.) infusion during 40 min. Subjects had two inpatient treatment visits with a washout period of 21 days. The MADRS was conducted according to the MADRS structured interview guide (MADRS-SIGMA<sup>17</sup>) by trained physicians at pre-dose, 100 min, 24 h and 1 week after start of infusion during both treatment visits and at 2 weeks after start of infusion during the first visit only. A more detailed description of the study design will be available elsewhere (K. Recourt, G.E. Jacobs, N. Drenth, J. van de Grond, K. Nishigori, J.M.A. van Gerven, unpublished data).

### Reference IRT model application

Carmody et al.<sup>18</sup> reported two IRT models with different sets of item parameters based on MADRS questionnaire

responses in two different populations (Table S1). The first population consisted of 208 patients with MDD (89%) and 25 patients (11%) with bipolar disorder, suffering from treatment-resistant depression and participating in a study on adjunctive vagus nerve stimulation on top of their current medication (referred to as Carmody #1). The second population consisted of 985 patients with MDD, who participated in a randomized, placebo- and comparator-controlled clinical trial testing a new antidepressant. In this second population, treatment-resistant patients were excluded (Carmody #2). Population demographics are listed in Table S2. Data at study exit were used for IRT model development for both populations. Of these two populations, the first population is considered to be most alike the ketamine study population, because both studies included treatment-resistant patients, and therefore IRT model parameters from the first population are used for this analysis. IRT model parameters from the second population are used to investigate differences in results when a different, less comparable, reference population is used.

## Distribution of the data

IRT model parameters as reported by Carmody et al.<sup>18</sup> were used to extract values of  $\psi_{kij}$ , where  $k$  is treatment visit (ketamine or placebo),  $i$  is subject, and  $j$  is scheduled time.  $\psi_{kij}$  was determined by finding the  $\psi$  value for which the likelihood to observe the provided set of item scores was largest, given the (fixed) ICC curves. More information about the IRT model structure and ICC curves is provided in the Appendix S1.

Three different approaches regarding the assumed distribution of the data were tested while the item parameters remained fixed. In approach A, the latent variable was not assumed to follow any prespecified distribution. Separate parameters ( $\theta$ ) were assigned to each measurement of each subject during each treatment and were all simultaneously estimated:

$$\psi_{kij} = \theta_{kij} \quad (1)$$

In approach B, it was assumed that all measurements at a certain timepoint during one specific visit arose from the same normal distribution:

$$\psi_{kij} = \theta_{kj} + \eta_{kij} \quad (2)$$

$$\eta_{kij} \sim N(0, \omega_{kj}^2) \quad (3)$$

where  $\theta$  is the estimated population mean per treatment per timepoint and  $\eta$  is the individual variability for this measurement resulting from a normal distribution with variance  $\omega^2$ ,

which is estimated per treatment per timepoint. In approach C, no parameters were estimated as it was assumed that all data belonged to the same distribution of the reference population (i.e., a standard normal distribution; a common practice in IRT model development to ensure the model parameters are identifiable<sup>3</sup>):

$$\psi_{kij} = 0 + \eta_{kij} \quad (4)$$

$$\eta_{kij} \sim N(0, 1) \quad (5)$$

where  $\eta_{kij}$  was resampled from the standard normal distribution for each timepoint per treatment per subject.

$\psi_{kij}$  were derived using the final parameter estimates in case of approach A and the Empirical Bayes Estimates (EBEs) in case of approaches B and C.

## Reference population

To assess the possible bias related to a wrong choice of model, the second IRT model reported by Carmody et al.<sup>18</sup> was used to determine values of  $\psi_{kij}$ . Approach C was used for the distribution of the data (Equations 4 and 5) and results of this approach were thus further specified to Carmody #1 or Carmody #2, respectively.

## Model evaluation

The ICCs of the two IRT models from Carmody et al.<sup>18</sup> were visualized and the ability of the models to describe the data was evaluated with goodness-of-fit (GOF) plots.<sup>3</sup> Frequency-plots (also known as mirror plots) were created by simulating the individual item scores 1000 times using the population parameters (approach A) or EBEs (approaches B and C) that were derived from fitting the models as values for  $\psi$ . For each determined  $\psi$  value, the cumulative probabilities for the increasing scores were calculated per item using the ICC's and a score was then simulated by drawing a random number from a uniform distribution between 0 and 1. The median and 95% prediction interval of the relative frequencies of the simulated scores per item were then plotted over the observed responses. Additionally, generalized additive models (GAMs) were used to create non-parametric ICC smooth plots of the ICCs based on both observed and simulated data. Uncertainty in estimation of population parameters (approaches A and B) and determination of EBEs (approaches B and C) was assessed with their relative standard error (RSE) obtained from the NONMEM reported variance-covariance matrix and variance of the posterior density distribution, respectively.

## Linear mixed effects modeling

To assess the impact of the different approaches and to determine the treatment effect, linear mixed effects models were fitted to (1) the total score data and (2) the  $\psi_{kij}$  data derived from each approach (A, B, C Carmody #1 or C Carmody #2). The linear mixed effects model structure was prespecified, based on analysis methods from the original statistical analysis plan and fixed factors consisted of treatment, visit, time, and treatment by time, and random factors consisted of subject, subject by treatment, and subject by time and the predose value as covariate. Influence on determination of  $\psi_{kij}$  value and treatment effect size due to chosen approach (A, B or C Carmody #1 or #2) was investigated by inclusion of approach as fixed factor and inclusion of an interaction term between treatment and approach, respectively. Inclusion of these terms and the treatment effect in general were considered significant when  $p < 0.05$  as resulting from the analysis of covariance, using the restricted maximum likelihood method. A more detailed description of the analysis is available in Appendix S1.

To visually compare the difference in statistical power resulting from previously described approaches, the standard error of the contrast was also used to calculate the power to determine treatment effect in a cross-over study for different samples sizes with a one sample  $t$ -test.

A schematic representation of the different steps described in the methods section is shown in Figure S1.

## Software

The  $\psi_{kij}$  values were obtained using maximum likelihood approximation in NONMEM (version 7.5) with the

Laplacian estimation method.<sup>19</sup> Data transformation, analysis, and visualization was done in R (version 4.0.3). Automatic generation of model code and GAM smooth plots was done using the Piraid package in R.<sup>20,21</sup> Linear mixed effects modeling was done in R using the lme4 and lmerTest packages.<sup>22,23</sup>

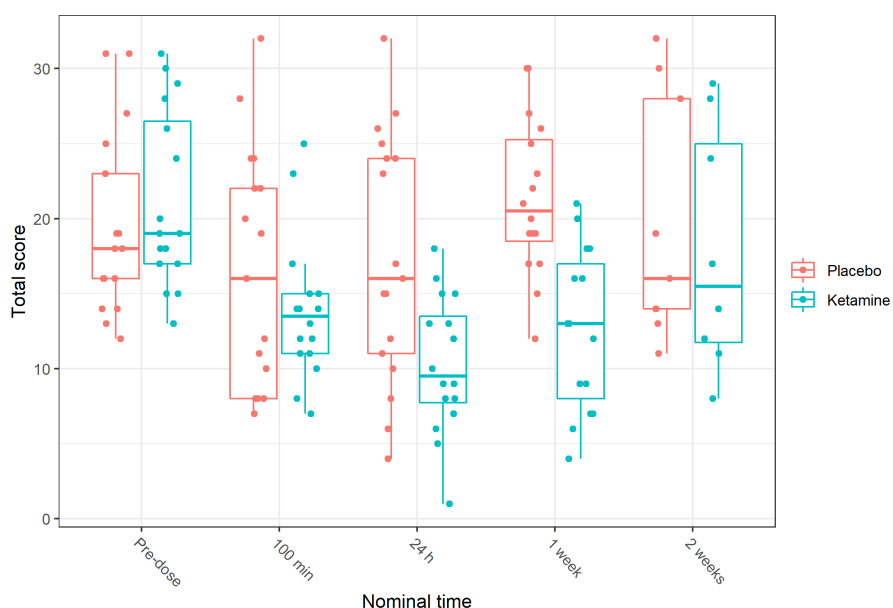
## RESULTS

### Clinical trial data

In total, 147 MADRS questionnaires were completed in 17 patients. One patient did not complete the study, of whom only data of the placebo treatment visit were available. All patients used concomitant antidepressant medication consistent with the selection criteria. Population demographics at study start are listed in Table S2. Total scores are visualized in Figure 1, showing that, compared to the data after placebo treatment, total score data after ketamine treatment were generally observed at a lower level at 100 min, 24 h, and 1 week after treatment. Item-specific and total score data are also available in the Tables S3 and S4.

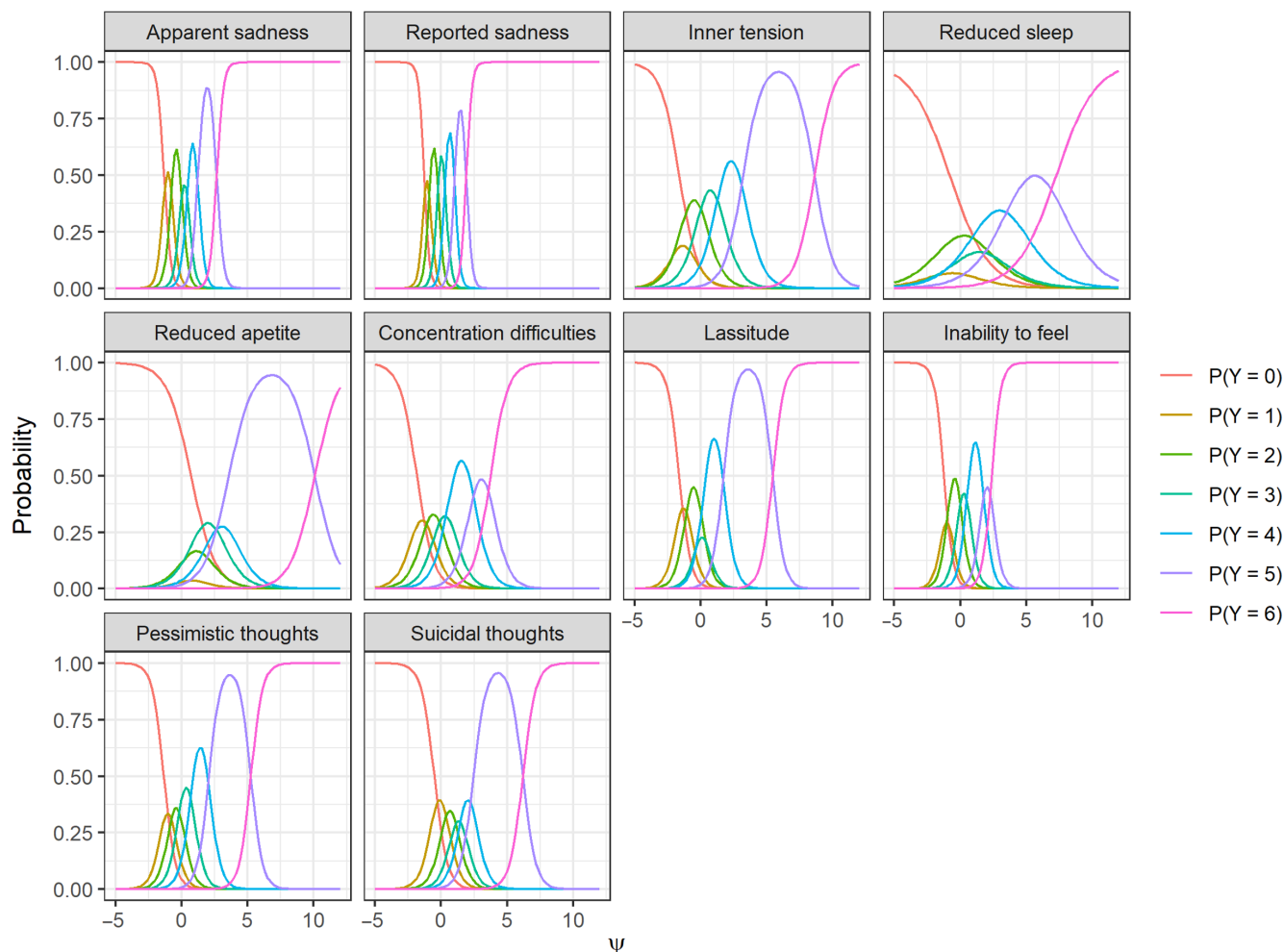
### Reference IRT model application and evaluation

ICC's of both available literature models are visualized in Figures 2 and 3. These figures show that the response patterns for item 1 (apparent sadness), 2 (reported sadness), and 8 (inability to feel) are highly similar between populations, whereas the curves for item 3 (inner tension) and 4



**FIGURE 1** Total scores of the MADRS over time visualized with boxplots for both placebo and ketamine treatment visits. Boxplots show the median, interquartile range (IQR) or 25th–75th percentiles (hinges) and the minimal and maximal observations or up to 1.5\*IQR (whiskers). MADRS, Montgomery-Åsberg Depression Rating Scale.





**FIGURE 2** Item characteristic curves of the graded response IRT model developed by Carmody et al.<sup>18</sup> based on the treatment-resistant MDD or bipolar disorder population (Carmody #1).  $P(Y=X)$ : probability of a certain score (0–6) to be given as response (Y) for a specific item. IRT, Item Response Theory; MDD, major depressive disorder.

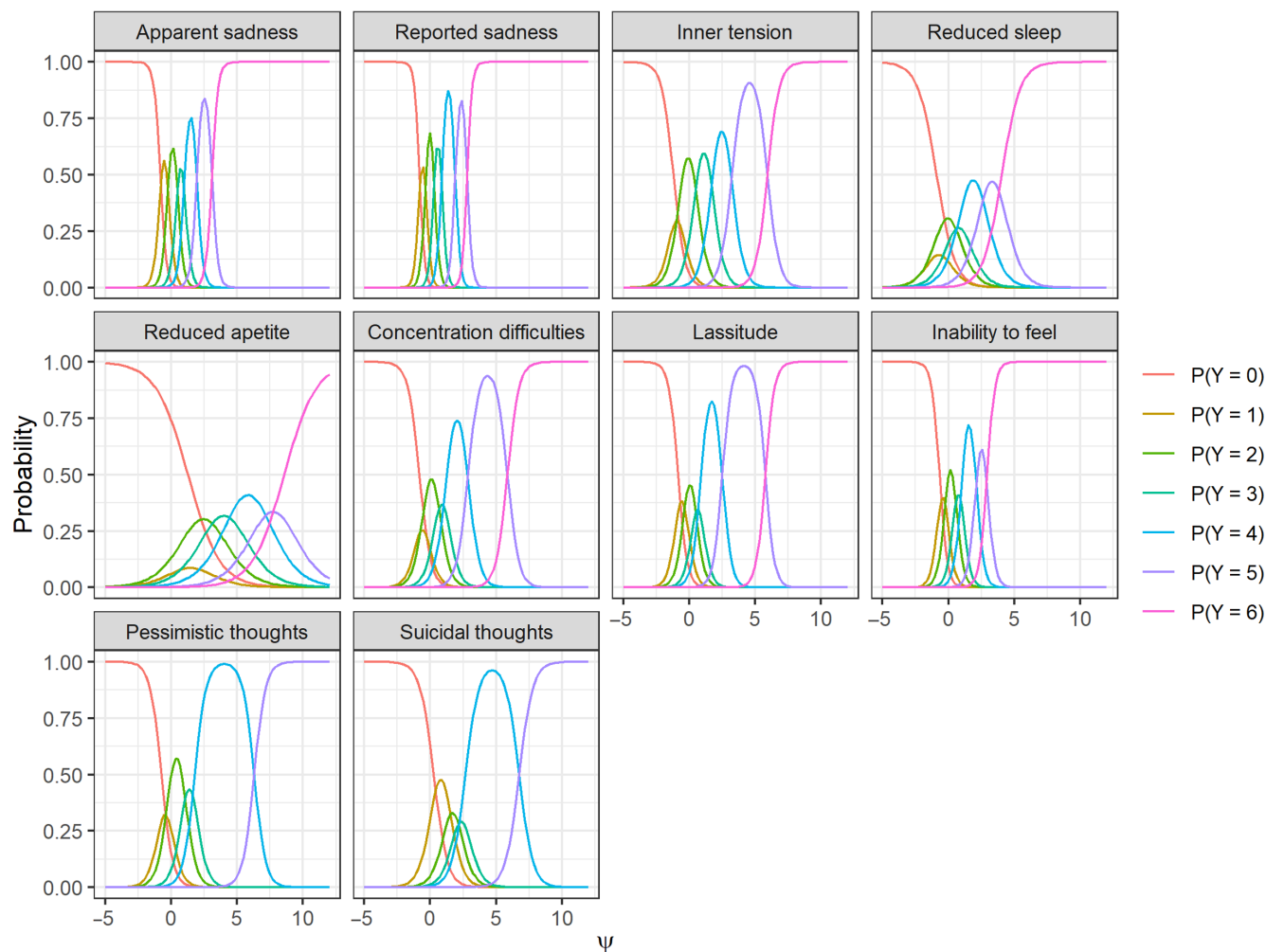
(reduced sleep) are spread over a smaller range of  $\psi$  values for the second population. In addition, the curve for the highest score of item 10 (suicidal thoughts) was not described for the second population, resulting in a highest possible score of 5 and not 6.

Estimated population parameter values of approaches A and B are listed in Tables S5 and S6. For both approaches, RSE% values of parameters often succeeded the 50% threshold, with a median RSE% value of 39.7% (interquartile range [IQR]: 27.2%–68.3%) for approach A. For approaches B and C, median RSE% values of the EBEs were 68.0% (IQR: 47.0%–136.5%) and 39.7% (IQR: 27.6%–69.4%), respectively.

The individual  $\psi$  estimates for each approach are visualized in Figure 4. Regardless of the chosen approach, individual profiles showed highly similar trends over time and differences between approaches were very small compared to the observed range. One exception, however, is subject 14, where the different approaches resulted in more variable estimates during the ketamine treatment

visit at 24 h. Further inspection of the data showed that all items had been answered with 0 at this particular measurement, except for item 4 (reduced sleep) which was answered with 1. Use of different reference populations (approach C Carmody #1 vs. C Carmody #2) resulted in an overall increase in  $\psi$  values, whereas individual trends over time remained similar.

As no large differences between derived  $\psi$  values were observed except for the sensitivity regarding scores almost all being 0 and given that population parameter estimates resulting from approaches A and B were uncertain, further IRT model evaluation focused on approach C. The ability of the reference model to adequately describe the data was evaluated with frequency plots and GAM smooth plots. The low number of observations available resulted in a very discrete scale of frequencies, which should be regarded with care when compared to the predicted probabilities. Especially in case of the GAM smooth plot (Figures S2 and S3), from which it only can be concluded



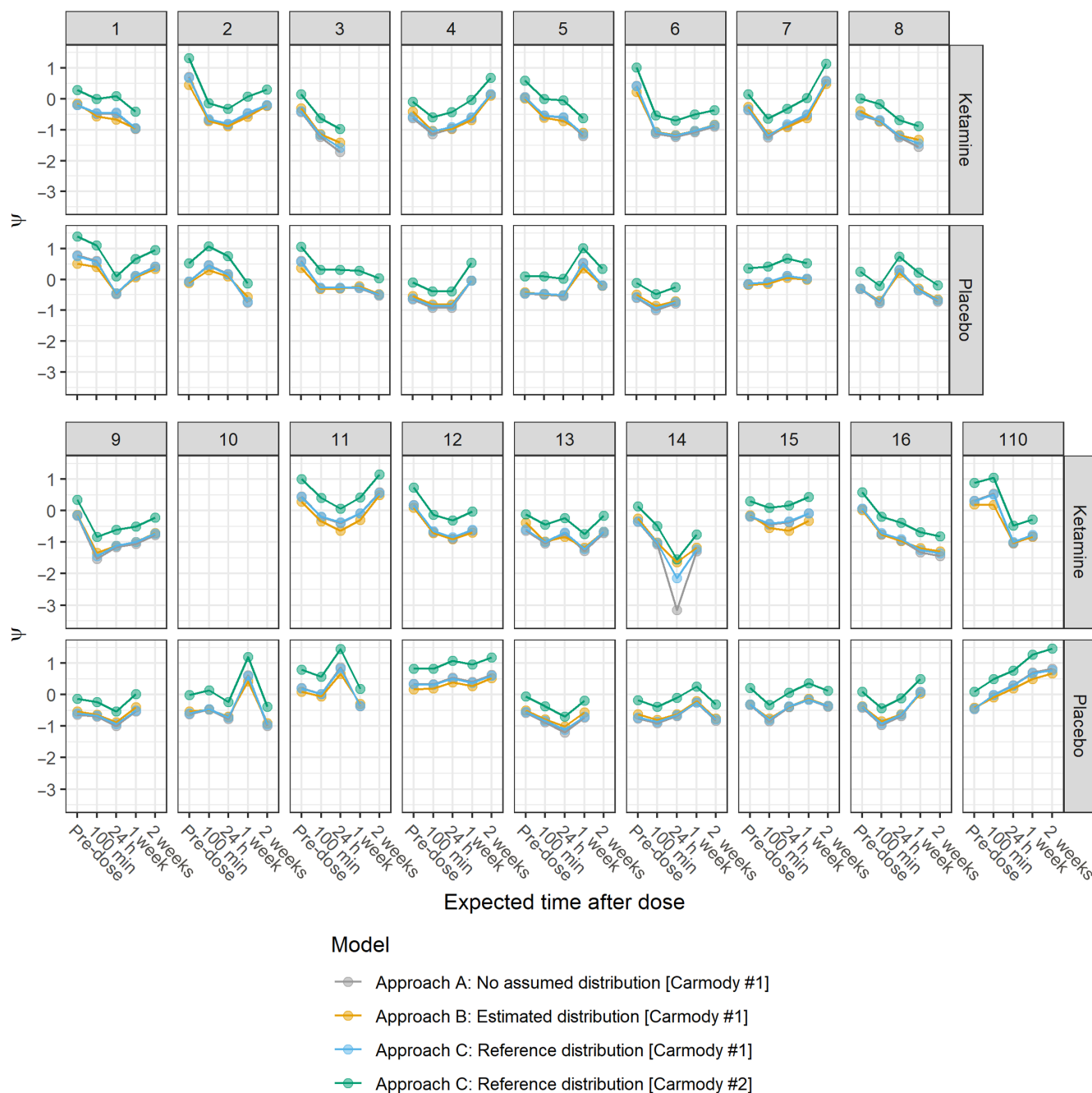
**FIGURE 3** Item characteristic curves of the graded response IRT model developed by Carmody et al.<sup>18</sup> based on the non-treatment-resistant MDD population (Carmody #2).  $P(Y=X)$ : probability of a certain score (0–6) to be given as response ( $Y$ ) for a specific item. IRT, Item Response Theory; MDD, major depressive disorder.

that not enough data are available for it to be useful for evaluation.

The frequency plot (Figure 5) shows that for most items the observed relative frequency for a certain score is within the 95% prediction interval of the simulated probability for that score. In general, the probability for the lowest score ( $Y=0$ ) is mispredicted most often, which can be seen by the discrepancy in height of the observed and simulated bars, especially for items 4 (reduced sleep), 5 (reduced appetite), and 10 (suicidal thoughts). Higher scores were sometimes also observed outside the simulated 95% prediction interval, such as the scores of 3 for item 3 (inner tension) and 3 and 4 for item 6 (concentration difficulties). Only minor differences in simulated relative frequencies between the two different reference populations (Carmody #1 and Carmody #2) can be observed, as the 95% prediction intervals overlap for most item scores with the exceptions of score 0 for item 4 (reduced sleep) and score 2 for item 9 (pessimistic thoughts).

## Linear mixed effects modeling

Two linear mixed effects models were fitted, describing either the total score data or the  $\psi$  values resulting from the different approaches (A, B, C Carmody #1 or C Carmody #2). At first, inclusion of approach as fixed factor in the linear model describing  $\psi$  was significant ( $p < 0.001$ ), showing that significant differences in  $\psi$  values exist between approaches. Refitting of the model with only  $\psi$  data resulting from approaches A, B, and C Carmody #1 to compare between assumptions on distribution only, resulted in no significance for approach as fixed factor ( $p = 0.61$ ), whereas refitting with  $\psi$  data from approaches C Carmody #1 and C Carmody #2 to compare between reference models only, did result in significance ( $p < 0.001$ ), which is in line with the results previously shown in Figure 4. Treatment effect size was not affected by choice of approach nor reference population, because inclusion of the interaction term between



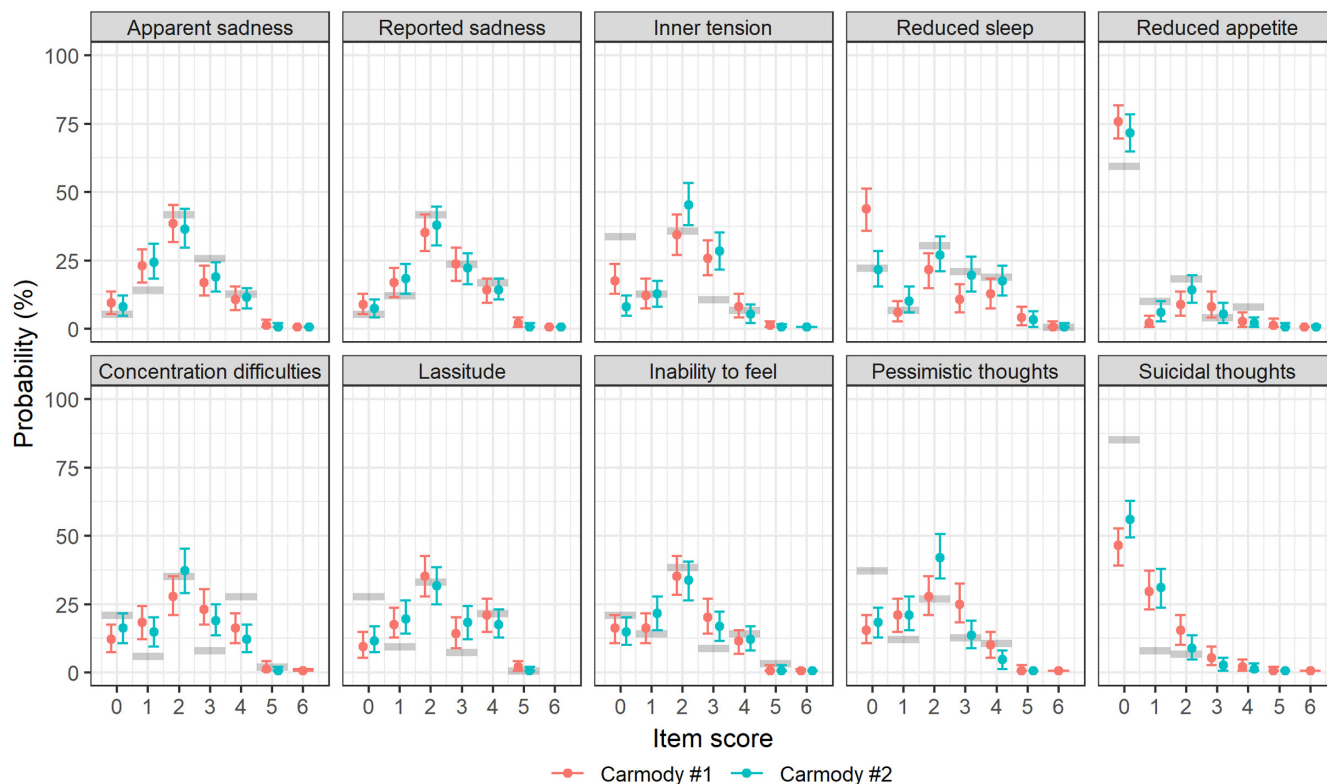
**FIGURE 4** Derived  $\psi_{kij}$  values over time per subject per treatment using item parameters reported by Carmody et al.<sup>18</sup> based on different assumptions for the distribution of the data. Item parameters were either based on patients with treatment resistant MDD or bipolar disorder (Carmody #1) or non-treatment resistant MDD (Carmody #2). Approach A–C are described in the Methods section.  $\psi$  = latent variable (depression severity). MDD, major depressive disorder.

treatment and approach was not significant in any of the models discussed above (all approaches included:  $p=0.88$ ; approaches A, B, and C Carmody #1 included:  $p=0.79$ , approaches C Carmody #1 and C Carmody #2 included:  $p=0.97$ ).

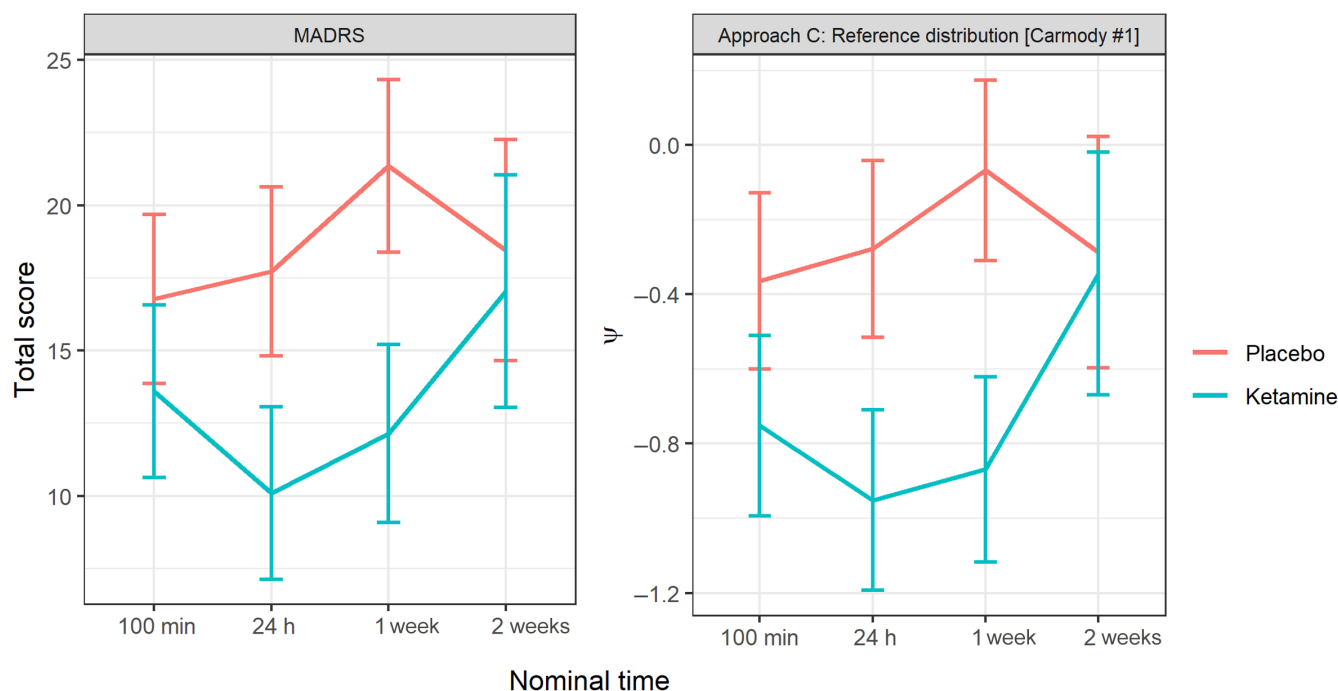
The same linear mixed effects model structure as was used for the total score was used to fit the  $\psi$  data resulting from approach C Carmody #1. Resulting parameters of both linear mixed effects models are listed in Table S7.

Estimated means of the contrast between placebo and ketamine treatment were 5.355 (95% confidence interval [CI]: 2.571–8.139,  $p=0.0009$ ) for the total score and 0.480 (95% CI: 0.284–0.676,  $p<0.0001$ ) for  $\psi$ . Treatment effect of ketamine on the total MADRS score was significant, but use of  $\psi$  resulted in a lower  $p$  value. Nevertheless, the general trends in estimated mean  $\psi$  and total score over time, shown in Figure 6, are highly similar. The difference in sample size required to determine a treatment effect in a

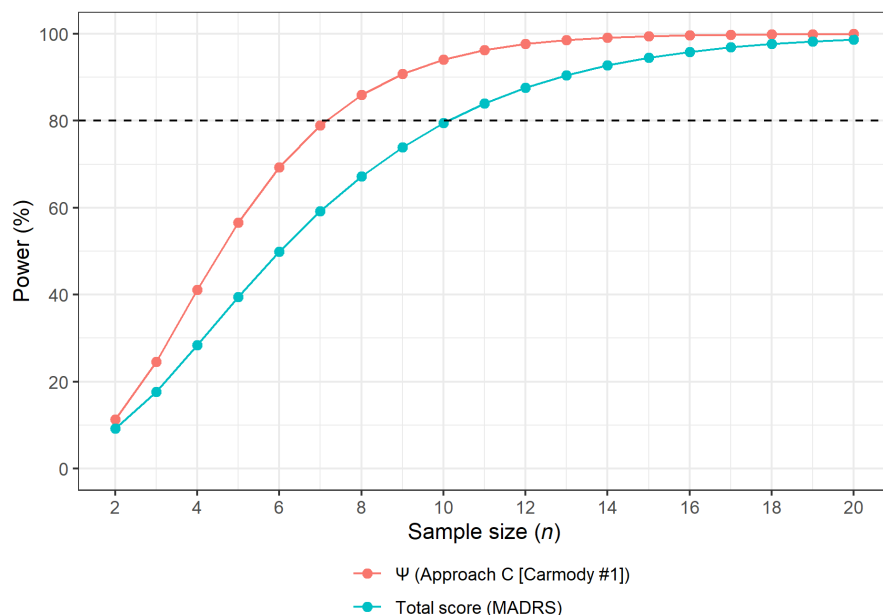




**FIGURE 5** Frequency plots of the probability for an observed or simulated score as predicted using the IRT model parameters reported by Carmody et al.<sup>18</sup> while using approach C for the distribution of the data. Simulated (points and error bars) and observed (gray lines) scores are stratified per item. Item parameters were based on patients with treatment-resistant MDD or bipolar disorder (Carmody #1) or patients with non-treatment resistant MDD (Carmody #2). The simulated scores show the median and 95% prediction intervals (error bars) of 1000 simulations. IRT, Item Response Theory; MDD, major depressive disorder.



**FIGURE 6** Estimated means over time per treatment resulting from the linear mixed effects modeling of total score or  $\psi$  data. Error bars represent the 95% confidence interval of the estimated mean.  $\psi$  = latent variable (depression severity). MADRS, Montgomery-Åsberg Depression Rating Scale.



**FIGURE 7** Calculated power to determine a treatment effect of ketamine on  $\psi$  or total score over different sample sizes in a cross-over design. MADRS, Montgomery-Åsberg Depression Rating Scale.

cross-over study when using total score versus  $\psi$  is further demonstrated in Figure 7.

## DISCUSSION

This paper reports the application of a reference IRT model to transform separate MADRS item scores derived from a clinical study to a latent variable,  $\psi$ , which can be considered a value for depression severity. In this randomized, double-blind, placebo-controlled, cross-over study, 17 patients with MDD received placebo and a single i.v. administration of an established antidepressant, ketamine. The transformed data were used to evaluate treatment effect and compared to the more common analysis of total MADRS score. To assess potential biases related to the use of a reference IRT model, different assumptions regarding the prespecified distribution of  $\psi$  data were tested and two reference IRT models were used. It was shown that assuming the data to be part of the reference distribution is preferred over estimating distribution parameters, as these cannot be estimated with high certainty while resulting  $\psi$  values are similar. The individual trends over time were generally unaffected by choice of assumption, except for data close to the minimum and maximum values (i.e., [almost] all item responses are either 0 or 6). Analysis of the data with a linear mixed model indicated that resulting  $\psi$  values were dependent on reference population, although treatment effect sizes were similar between both groups. When compared to the use of total MADRS scores, the trend of  $\psi$  over time for both the ketamine and placebo treatments was similar. However, use of  $\psi$  derived from separate MADRS item scores as outcome in a linear mixed model analysis was shown to be superior in

terms of precision of the treatment contrast estimate over the whole study period, demonstrating the usefulness and successful application of IRT as analysis method in early PoC clinical trials when only limited data are available.

A reference IRT model can only be applied when one is confident that the reference population and study population have similar psychometric characteristics, as it can then be assumed that the previously found relation between disease severity and response can indeed be used for extrapolation. The studied population here and the first reference population reported by Carmody et al.<sup>18</sup> both suffered from (relatively) treatment resistant MDD and were therefore labeled as most compatible, but this reference population also included a small number of patients with bipolar depression. Furthermore, the reference population was relatively older (mean: 47.2 years, SD: 8.9 vs. mean: 26.8, SD: 7.3) and had higher HAM-D-17 (mean: 21.9, SD: 4.4 vs. mean: 19.2, SD: 1.9) and MADRS (mean: 31.9, SD: 6.7 vs. mean: 26.6, SD: 4.0) total scores at the start of the study compared to the study population prior to dosing. Although it could be argued that these differences in demographics are not likely to influence the way patients respond to a rater-based interview, this cannot be ruled out and should be considered when interpreting the data. For example, cultural differences may result in different “true” discrimination or difficulty parameters for certain items for the study population compared to the reference population.

When results based on the first reference population were compared with the second reference population, which did not include any treatment-resistant patients with MDD and would thus theoretically pose larger concern regarding compatibility, individual trends over time barely differed and the treatment effect did not change,

indicating similar psychometric properties of both populations. Interestingly, however, use of the second reference model did result in an overall increase in  $\psi$ . Because the scale of  $\psi$  is in SD to the mean (i.e., 0) of the reference population, the distance in  $\psi$  of the study population to the mean depression severity of the reference population was increased for the second population compared to the first, confirming that the second population was generally less severely depressed than the first (reflected by lower HAM-D-17 and MADRS scores at baseline as well). Drawing conclusions based on absolute  $\psi$  values thus may be questionable when reference IRT models were used to obtain them, considering the large impact of choice of reference population, and research should focus on relative changes in  $\psi$  over time or between groups instead. Of note, the clinical relevance of changes in MADRS total scores are often assessed in terms of “remission” (total score  $\leq 12$ ) and “response” ( $\geq 50\%$  reduction).<sup>24</sup> In light of results presented here, a similar target for “response” (i.e., a relative reduction) might be defined on the  $\psi$  scale, but the basic definition of “remission” (an absolute decrease) based on a threshold value would need to be evaluated.

The ability of a model to correctly describe or predict a new dataset is usually readily evaluated by assessment of parameter uncertainty and GOF plots, which was also performed in this report. As both approaches for which parameters had to be estimated (i.e., A and B) resulted in a large number of parameters with RSE greater than 50%, which is not unexpected considering the limited amount of data, one may question how certain we are of the  $\psi$  output. Indeed, in case of the outlier included in this dataset, it is possible these approaches either overestimated (approach A) or underestimated (B) the true value. Therefore, assuming the data are part of the reference distribution (approach C) would be preferred for using IRT model outcomes for further statistical analysis in early phase clinical trials with small datasets, considering this approach avoids imprecise parameter estimation and is affected less by outliers.

Evaluation of the IRT model performance using GOF plots commonly used for IRT model development,<sup>3</sup> on the other hand, did not provide very valuable insights as the small size of the dataset and its discrete nature made it hard to interpret results. This effectively rendered the GAM smooth plots uninformative whereas they would normally provide more capability of identifying ICC misspecification than the frequency plots. Because no longitudinal or disease progression model was developed, evaluation through commonly used visual predictive checks by simulation was not possible. Further stratification of the frequency plot per timepoint might have provided comparable information, however, this would have resulted in even more discreteness of the data and interpretation would be hampered

further. The frequency plot showed that the model was generally able to predict the data accurately, but mispredictions did occur throughout the range of scores and especially for the lowest score for some items. However, the discrimination parameters of most of these items were low, which is indicative for the fact that the items do not contribute much information for the determination of the latent variable.<sup>18</sup> Therefore, the impact of a possible model misspecification for these items is small.

Because other IRT models are lacking for this specific questionnaire, utilization of the IRT model of Carmody et al.<sup>18</sup> was the only option for the IRT-based analysis of the MADRS presented here. The sample size of the population upon which Carmody #1 was based was sufficient but would preferably have been larger ( $N=233$  vs. recommended minimum of  $\sim 250$ <sup>3,16</sup>). Still, results using Carmody #2 ( $N=985$ ) were similar. Estimation of IRT model parameters with the available clinical dataset ( $N=147$ ), while using the values reported by Carmody et al. as prior information might have been an interesting alternative. However, this would have required merging of scores (e.g., to categories of 0, 1, 2, 3, and  $>4$ ), because there still would not have been enough observations of some of the higher item scores to estimate their respective difficulty parameters (see Table S3), thus resulting in a loss of information still. As correct use of reference model is an important assumption, this issue of reference model availability and quality needs to be addressed to further enhance use of IRT-based analysis method for early phase clinical trials. For example, one might suggest an international collaborative initiative to gather enough data to build robust reference IRT models, not only for the MADRS, but also for other questionnaires commonly used for clinical drug development in other therapeutic areas.

To conclude, higher sensitivity to detect treatment effects in early PoC clinical trials in psychiatry can potentially be achieved if reference IRT models are (also) applied in the analysis of rater-based questionnaire data instead of total scores, as we demonstrated using data from a study evaluating the effects of single dose ketamine in MDD with the MADRS as the primary outcome as case study. The weight carried by the assumptions regarding the selected reference population and data distribution, when robust methods for evaluation of IRT model performance for small datasets are still lacking, might hamper application of reference IRT models as primary analysis of rater-based questionnaire data in early phase clinical trials. Nevertheless, if IRT-based analyses were to be used systematically for exploratory purposes, as is often done at this stage in drug development, it shows high potential to provide further valuable insights regarding the determination of possible treatment effects in subsequent development phases of novel compounds.

## AUTHOR CONTRIBUTIONS

M.E.O., K.R.B., M.L.d.K., K.R., G.E.J., and M.J.v.E. wrote the manuscript. M.E.O., K.R.B., M.L.d.K., and M.J.v.E. designed the research. M.E.O. performed the research and analyzed the data.

## ACKNOWLEDGMENTS

The authors would like to thank Sumitomo Pharma Co., Ltd. for their permission to use the data, and for their continuous interest and support of this analysis.

## FUNDING INFORMATION

Funding for this study was provided by Sumitomo Pharma Co., Ltd.

## CONFLICT OF INTEREST STATEMENT

The authors declared no conflicts of interest for this work.

## ORCID

Marije E. Otto  <https://orcid.org/0000-0002-5767-604X>  
 Michiel J. van Esdonk  <https://orcid.org/0000-0001-8159-0273>

## REFERENCES

1. FDA. [Draft guidance] Major depressive disorder: developing drugs for treatment guidance for industry. 2018.
2. Montgomery A, Åsberg M. Scale designed to be sensitive to change. *Br J Psychiatry*. 1979;134:382-389.
3. Ueckert S. Modeling composite assessment data using item response theory. *CPT Pharmacometrics Syst Pharmacol*. 2018;7:205-218.
4. Baker FB. *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation; 2001.
5. Bock RD. A brief history of item response theory. *Educ Meas Issues Pract*. 1997;16:21-33.
6. Gortler R, Fox JP, Twisk JWR. Why item response theory should be used for longitudinal questionnaire data analysis in medical research. *BMC Med Res Methodol*. 2015;15:1-12.
7. Soland J. Evidence that selecting an appropriate item response theory-based approach to scoring surveys can help avoid biased treatment effect estimates. *Educ Psychol Meas*. 2021;82:376-403. doi:10.1177/00131644211007551
8. Ueckert S, Plan EL, Ito K, et al. Improved utilization of ADAS-cog assessment data through item response theory based pharmacometric modeling. *Pharm Res*. 2014;31:2152-2165.
9. Llanos-Paez C, Ambery C, Yang S, et al. Improved decision-making confidence using item-based pharmacometric model: illustration with a phase II placebo-controlled trial. *AAPS J*. 2021;23:79.
10. Chen C, Jönsson S, Yang S, Plan EL, Karlsson MO. Detecting placebo and drug effects on Parkinson's disease symptoms by longitudinal item-score models. *CPT Pharmacometrics Syst Pharmacol*. 2021;10:309-317.
11. Chae D, Chung SJ, Lee PH, Park K. Predicting the longitudinal changes of levodopa dose requirements in Parkinson's disease using item response theory assessment of real-world unified Parkinson's disease rating scale. *CPT Pharmacometrics Syst Pharmacol*. 2021;10:611-621.
12. Cerou M, Peigné S, Comets E, Chenel M. Application of item response theory to model disease progression and Agomelatine effect in patients with major depressive disorder. *AAPS J*. 2020;22:1-13.
13. Buatois S, Retout S, Frey N, Ueckert S. Item response theory as an efficient tool to describe a heterogeneous clinical rating scale in de novo idiopathic Parkinson's disease patients. *Pharm Res*. 2017;34:2109-2118.
14. Guk J, Chae D, Son H, Yoo J, Heo JH, Park K. Model-based assessment of the benefits and risks of recombinant tissue plasminogen activator treatment in acute ischaemic stroke. *Br J Clin Pharmacol*. 2018;84:2586-2599.
15. Houts CR, Morlock R, Blum SI, Edwards MC, Wirth RJ. Scale development with small samples: a new application of longitudinal item response theory. *Qual Life Res*. 2018;27:1721-1734.
16. Reise SP, Yu J. Parameter recovery in the graded response model using MULTILOG. *J Educ Meas*. 1990;27:133-144.
17. Williams JBW, Kobak KA. Development and reliability of a structured interview guide for the Montgomery-Åsberg depression rating scale (SIGMA). *Br J Psychiatry*. 2008;192:52-58.
18. Carmody TJ, Rush AJ, Bernstein I, et al. The Montgomery Åsberg and the Hamilton ratings of depression: a comparison of measures. *Eur Neuropsychopharmacol*. 2006;16:601-611.
19. Beal SL, Sheiner LB, Boeckman AJ. *NONMEM 7.5.0 User Guides*. ICON Development Solutions; 1989-2020.
20. Nordgren R, Ueckert S, Karlsson MO. piraid: An aid for development and diagnosis of pharmacometric 'IRT' models. R package version 0.4.1 2020.
21. R Core Team. *R: a language and environment for statistical computing*. 2020. <https://www.r-project.org/>
22. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: tests in linear mixed effects models. *J Stat Softw*. 2017;82:1-26.
23. Bates D, Maechler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Software*. 2015;67:1-48.
24. Fedgchin M, Trivedi M, Daly EJ, et al. Efficacy and safety of fixed-dose Esketamine nasal spray combined with a new Oral antidepressant in treatment-resistant depression: results of a randomized, double-blind, active-controlled study (TRANSFORM-1). *Int J Neuropsychopharmacol*. 2019;22:616-630.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Otto ME, Bergmann KR, de Kam ML, Recourt K, Jacobs GE, van Esdonk MJ. Item response theory in early phase clinical trials: Utilization of a reference model to analyze the Montgomery-Åsberg Depression Rating Scale. *CPT Pharmacometrics Syst Pharmacol*. 2023;12:1425-1436. doi:[10.1002/psp4.13018](https://doi.org/10.1002/psp4.13018)